Supplement of

Brief communication: Towards a universal formula for the probability of tornadoes

Roberto Ingrosso et al.

Correspondence to: Piero Lionello (piero.lionello@unisalento.it)

The copyright of individual parts of the supplement might differ from the article licence.
1 Methods

1.1 Definitions of meteorological variables

We use the following definitions of the meteorological variable WMAX (m/s), WS700 (m/s), SRH900 (m$^2$/s$^2$), LCL (m):

$$WMAX = \sqrt{2 \cdot CAPE}.$$  \hfill (S1)

where

$$CAPE = \int_{z_{el}}^{z_{lfc}} B \, dz,$$  \hfill (S2)

and $z_{el}$, $z_{lfc}$, and $B$ are, respectively, the equilibrium level, the level of free convection and the Buoyancy force per unit mass,

$$WS_{700} = ||u_{700hPa} - u_{10m}||,$$  \hfill (S3)

that is the module of the difference between the 10 m and 700 hPa winds,

$$SRH_{900} = \int_{z_0}^{z(900)} \nabla \times u(z) \cdot (u(z) - u_c) \, dz$$  \hfill (S4)

that is the scalar product between the storm-relative motion and the vorticity of the horizontal wind, where $u$ and $u_c$ are the wind speed and the translation speed of the storm [1],

$$LCL \approx 125 \cdot (T_{2m} - T_{d2m}),$$  \hfill (S5)

where $T$ and $T_d$ are respectively the 2 m air and dew point temperature, by using the Espy’s approximation [2].

1.2 Univariate and bivariate regression models

Below, we briefly outline the regression models adopted in this work. Here, $y$ is the $\log_{10}$ of the conditional probability of tornado occurrence. In the univariate analysis, the linear regression

$$y = a + bx$$  \hfill (S6)

is used for $WS_{700}$, $SRH_{900}$, LCL. Instead, the non-linear regression

$$y = a + \frac{x}{b + cx}$$  \hfill (S7)

is used for WMAX.

For the bivariate probability, a non-linear regression model has been adopted for all the pairs $(x_1, x_2)$ involving WMAX:
\[ y = a + \frac{x_1}{b + c x_1/x_2} \]

(S8)

for (WMAX, WS700)—see Fig. 2 in the main text,

\[ y = a + x_1/b^{x_1/x_2} \]

(S9)

for (WMAX, LCL),

\[ y = a + b x_1^c |x_2|^d \]

(S10)

for (WMAX, SRH900), and the multiple linear model

\[ y = a + b x_1 + c x_2 \]

(S11)

for the remaining pairs.

### 1.3 Estimate of the conditional probability

Below, we briefly outline the statistical tools adopted in this work. Here, the main target is to provide an estimate of the probability of tornado occurrence *conditional on* the fact that the variable(s) considered take on values in a given range. In the single-driver case the procedure is as follows.

1. For each driver (say, \( X \)), the observed range spanned by the variable is first partitioned into \( K \) sub-intervals \( \Delta_i \)'s, with \( i = 1, \ldots, K \). The sub-intervals are chosen to be equiprobable, i.e. they contain the same number of observations of \( X \); the rationale (a least-informative strategy) is to use a common sample size for the calculation of the conditional probability in any given sub-interval. In turn, the intervals may not have the same length, since their boundaries are computed via the order statistics associated with the available sample, and thus the distance between successive empirical quantiles (and hence the distance between the boundaries of any given sub-interval) depends on the probability distribution \( F_X \) of \( X \), which is not the Uniform one.

2. In order to choose a reasonable number \( K \) of sub-intervals, several numerical experiments were carried out. As a result, the choice \( K = 17 \) represents a rational compromise between (i) a number of bins (predictors) sufficient for robust regressions (see, later, the details), and (ii) a number of observations and tornado occurrences in each sub-interval sufficient for a robust statistical analysis.

3. The estimate of the probability \( p^X_i \) of a tornado occurrence, conditional on the fact that the driver \( X \) takes on value in a given sub-interval \( \Delta_i \), with \( i = 1, \ldots, K \), is carried out according to the standard formula

\[ p^X_i = P(T \mid I_i) = \frac{P(T \cap I_i)}{P(I_i)} \]

(S12)
where \( T \) and \( I_i \) indicate, respectively, the events “a tornado has occurred” and “\( X \) belongs to the sub-interval \( \Delta_i \)”. The probabilities of interest are approximated as follows via Maximum Likelihood estimators.

First, the number of times \( N_i^X \) that \( X \) belongs to \( \Delta_i \) is computed, for \( i = 1, \ldots, K \). Then, the number \( N_i^T \) of tornadoes occurrences in \( \Delta_i \) is calculated. In turn, the conditional probability of interest is given by

\[
p_i^X = \frac{N_i^T}{N_i^X}, \quad i = 1, \ldots, K.
\]

The uncertainty affecting the estimates of the \( p_i^X \)’s is assessed via suitable Bootstrap (Monte Carlo) procedures [3].

The estimate of the \( p_i^X \)’s, for all the four variables WMAX, WS\(_{700}\), SRH\(_{900}\) and LCL, are plotted in Fig. 1 in the main text.

A crucial point of interest concerns the investigation of a possible relation between the \( p_i^X \)’s and the magnitude of the drivers \( X \)’s considered. For this purpose, both linear and non-linear regressions are carried out, where the \( p_i^X \)’s play the role of responses, and the median values of the sub-intervals \( \Delta_i \)’s, with \( i = 1, \ldots, K \), the role of predictors (see Eqs. S6 and S7).

Along the lines of the procedure adopted in the single-driver case, a regression approach is also used to provide a model for the probability of tornado occurrence conditional on the fact that two variables (say, \( X_1 \) and \( X_2 \)) take on value in a specific sub-domain of the corresponding \((x_1, x_2)\) plane. In turn, by analogy with the single-variable case, a \( K \times K \) (i.e., \( 17 \times 17 \)) grid matrix \( G \) is constructed, using as boundaries of the (rectangular) cells those of the marginal sub-intervals \( \Delta_i^{X_1} \)'s and \( \Delta_j^{X_2} \)'s, with \( i, j = 1, \ldots, K \). Then, the target is to estimate the probability \( p_{ij}^{X_1,X_2} \) of a tornado occurrence conditional on the fact that the pair \((X_1, X_2)\) takes on value in a given cell \( C_{ij} \) of \( G \). Note that no tornado occurrences were observed in some cells. Again, Eqs. S12–S13 can be used to provide Maximum Likelihood estimates of the conditional probabilities of interest.

In turn, six different bivariate analyses are carried out, considering one at a time the six pairs extracted from the four variables of interest. In particular, given the behavior of WMAX shown in the single-driver analysis, non-linear regressions are used for all the pairs involving WMAX, and linear regression for the remaining ones (see Eqs. S8–S11).

## 2 Geographical tornado distribution

A Gaussian kernel density estimation is provided to estimate a probability density function of the observed tornadoes over Europe and US during the period 2000-2018. (see Fig. S1). It can be seen that the peak of density over US (with a total of 2632 EF2+ tornadoes) is much higher than over EU (total of 441 EF2+ tornadoes). Peak positions are in northern Alabama and Mississippi, eastern Tennessee and
Kentucky, and southern Illinois for US, while the European peaks are in Germany, western Poland and eastern Czech Republic.

3 Slices of non-linear regressions

The fits of the conditional probability of tornado occurrence for the pairs of drivers \((WMAX, WS_{700})\), \((WMAX, SRH_{900})\) and \((WMAX, LCL)\) are carried out via non-linear regressions. In turn, in these cases, the traditional coefficient of determination \(R^2\) cannot provide a consistent measure of the goodness of the fits. In order to evaluate whether the fits are valuable, we cut the interpolating surface into 17 slices (see Figs. S2–S4), and compare the local profiles of the fits with the estimated conditional probabilities, for \(WMAX\) taking value on the whole range and increasing fixed values of the other variable. Apparently, in all cases the non-linear regression surfaces provide valuable fits of the behavior of the conditional probability over all the domain of interest.

4 Bivariate analysis

The additional figures regarding the bivariate analysis for pairs \((WMAX, SRH_{900})\), \((WMAX,LCL)\), \((WS_{700},LCL)\), \((WS_{700}, SRH_{900})\), and \((SRH_{900},LCL)\) are shown in Figs. S5, S6, S7, S8 and S9.

References


**Figures and Tables**

![EF2+ Tornado density](image)

**Figure S1:** Tornado density during the period 2000-2018 for Europe (top panel) and US (bottom panel). The density is calculated by means of a kernel density estimation over a grid of 0.5° horizontal resolution.
US+EU: $(X_1=W_{\text{MAX}}, X_2=W_{\text{S700}})$

Figure S2: Slices of the non-linear regression surface considering the pair $(W_{\text{MAX}}, W_{\text{S700}})$, as a function of $W_{\text{MAX}}$ for increasing values of $W_{\text{S700}}$. The markers indicate the estimated conditional probability of tornado occurrence, and the lines the corresponding fits.
US+EU: \((X_1=WMAX, X_2=SRH900)\)

**Figure S3:** Slices of the non-linear regression surface considering the pair \((WMAX, SRH900)\), as a function of \(WMAX\) for increasing values of \(SRH900\). The markers indicate the estimated conditional probability of tornado occurrence, and the lines the corresponding fits.
**Figure S4:** Slices of the non-linear regression surface considering the pair $(WMAX, LCL)$, as a function of $WMAX$ for increasing values of $LCL$. The markers indicate the estimated conditional probability of tornado occurrence, and the lines the corresponding fits.
Figure S5: Bivariate probability distribution for \((X_1 = WMAX, X_2 = SRH900)\). The coloured surface shows the empirical fit of \(y = \log_{10} P\). Upward/downward triangles represent empirical estimates located above/below the fitted surface. All values are reported according to the colour bar.

Figure S6: Bivariate probability distribution for \((X_1 = WMAX, X_2 = LCL)\). The coloured surface shows the empirical fit \(y = \log_{10} P\). Upward/downward triangles represent empirical estimates located above/below the empirical fit. All values according to the colour bar.
**Figure S7:** Bivariate probability distribution for $X_1 = WS_{700}$, $X_2 = LCL$. The coloured surface shows the empirical fit $y = \log_{10} P$. Upward/downward triangles represent empirical estimates located above/below the empirical fit. All values according to the colour bar.

**Figure S8:** Bivariate probability distribution for $X_1 = WS_{700}$, $X_2 = SRH_{900}$. The coloured surface shows the empirical fit $y = \log_{10} P$. Upward/downward triangles represent empirical estimates located above/below the empirical fit. All values according to the colour bar.
**Figure S9**: Bivariate probability distribution for $X_1 = SRH_{900}$, $X_2 = LCL$. The coloured surface shows the empirical fit $y = \log_{10}P$. Upward/downward triangles represent empirical estimates located above/below the empirical fit. All values according to the colour bar.